ONTOLOGY-BASED FRAMEWORK FOR DOCUMENT INDEXING

Djida Bahloul, Youssef Amghar, Pierre Maret
INSA de Lyon – LIRIS – FRE 2672. 7 avenue Jean Capelle. 69621, Villeurbanne – France.

Keywords: Semantic indexing, Ontology, Knowledge modelling

Abstract: The work presented in this paper addresses a project of the Computer Centre CIRTIL who supported it. This company wants to save and capitalize its knowledge and its know-how about the production activities, especially concerning the technical incidents relating to software applications encountered during the exploitation time. Indeed using a well accessing documents base, actors will be able to better solve problems. Our purpose is to focus on ontology-based framework for indexing documents. The domain ontology *OntoCIRTIL* has a structure which supports a semantic model based on semantic links and inference mechanisms. In this paper, we present a new model called S^3 which permits to model knowledge in upstream and index documents (or formalized knowledge) in downstream. To illustrate partial results, this model is then applied to *OntoCIRTIL*.

1 INTRODUCTION

Professionals of library sciences (i.e. librarians) accomplish indexing tasks in using indexing resources as a controlled vocabulary (thesaurus). The main characteristic of indexing resources resides in the ability to offer a way to carry out a relevant retrieval of document contents. Thus, it contains descriptors that are the keywords describing objects and their inter-relationships.

In this paper, we propose an indexing approach, based on ontologies, and which is more efficient than a simple use of taxonomy of concepts. Therefore, we build a domain ontology which has a significant capacity of expression thanks to the possibility of introducing semantic links, structural links and subsumption links. Compared to descriptors, concepts of ontology, are used first, to represent knowledge and then, to query the system in order to access and retrieve these knowledge.

We discuss the role of ontologies on the corporate memory building in section 2. Section 3 presents the environment of our work. Section 4 gives the description of our model called S^3 for representing and indexing knowledge. This model is applied to our domain ontology *OntoCIRTIL*.

2 ONTOLOGY, KNOWLEDGE REPRESENTATION

Knowledge is a combination of information and contexts required to perform a task by selecting, interpreting and evaluating information according on context of use (Weggeman, 1996). Ontologies are used as a coherent support to describe and to share knowledge. *“Ontologies open the way to move from a document oriented view of knowledge management to a content-oriented view, where knowledge items are interlinked, combined, and used.»* (Staab & al., 2001). Projects SHOE (Heflin & all, 2000) and Ontobroker (Benjamins & all, 1998) use ontologies to improve the searching abilities on the web. Both systems are logical reasoning based on ontological definitions. In SHOE ontologies are taxonomy hierarchies queried by users. A web page on the Internet can reference any ontology and exploit definitions using tags. Ontologies are used more and more in Knowledge Management System development (Van Heijst & all, 1997). They improve the knowledge engineering process.

2.1 Annotation with ontology

Ontologies can be used to improve information retrieval through annotations of the resources constituting a corporate memory (CM). Most recent

Bahloul D., Amghar Y. and Maret P. (2004).
ONTOLOGY-BASED FRAMEWORK FOR DOCUMENT INDEXING.
In *Proceedings of the Sixth International Conference on Enterprise Information Systems*, pages 269-274
DOI: 10.5220/0002606302690274

approaches employ ontologies as domain-specific vocabularies and concept structures according to explicitly specified conceptualisations. This approach extends traditional metadata technologies, such as taxonomies and thesauri, by additional relationships, axioms and general logical constraints to allow reasoning. The CoMMA project (Gandon & al, 2002) offers a solution to implement a CM based on ontologies and on agent technology. It promotes a wide vision of the document retrieval issue. The CM is composed of heterogeneous evolving documents, structured using semantic annotations expressed with concepts and relationships provided by a shared ontology. The approach for ontology-based knowledge management (Staab & al., 2001) includes a tool suite and a methodology for developing knowledge management systems. OntoAnnotate allows users to create objects and describe them with their attributes and relationships.

2.2 Ontology as an indexing resource

Thesaurus and ontology provide common properties such as the organization of terminology for covering of a broad range of terminology used in a particular domain, the use of hierarchical structure (terms are grouped into categories and subcategories). According to (Saadani & al, 2000) the more important differences is the informality and ambiguity of relationships in a thesaurus. Ontology introduces a host of structural and conceptual relationships including superclass/subclass/instance relationships, property values time relationships, and others depending on the representation language. Generally, ontology contains far more relationships, which are formally defined and unambiguous. The ontology can reason about the meaning of concepts by comparing logical concept structures. An example is given in (Desmontils & al, 2002) in indexing web sites with a terminology-based ontology. When concept C2 satisfies the requirement of being a specialization of concept C1, then C2 can automatically be classified below C1. This gives rise to query processing and searching which is not possible with a thesaurus.

2.3 Domain ontology: *OntoCIRTIL*

The domain we consider addresses the call centre activity of CIRTIL, a Computer Centre which has the following missions: i) management of the Information System of the Covering Branch; ii) setting in production and exploitation of software applications; iii) realization and/or contribution to the national and regional projects; iv) technical and functional help to customers. A first work allowed us to build ***OntoCIRTIL*** that intends to model technical and functional incidents, actors who treat these incidents, applications concerned by these incidents and the entities characterising these incidents, applications and actors. Main purposes of OntoCIRTIL are to represent knowledge, to permit indexing documents and consequently feeding CM and to reuse and share capitalized knowledge.

Several approaches to ontologies development have been proposed. To build *OntoCIRTIL*, we were inspired by methodologies proposed in (Uschold & all, 1996), (Guarino, 1995) and (Fernández-López & al., 1999). We chose a *middle-out* approach. Indeed, Uschold argues that this approach is most effective when the basic concepts in a domain are identified first (e.g., *Employee*), and later generalised (*Person*) and/or specialised (*Secretary*).

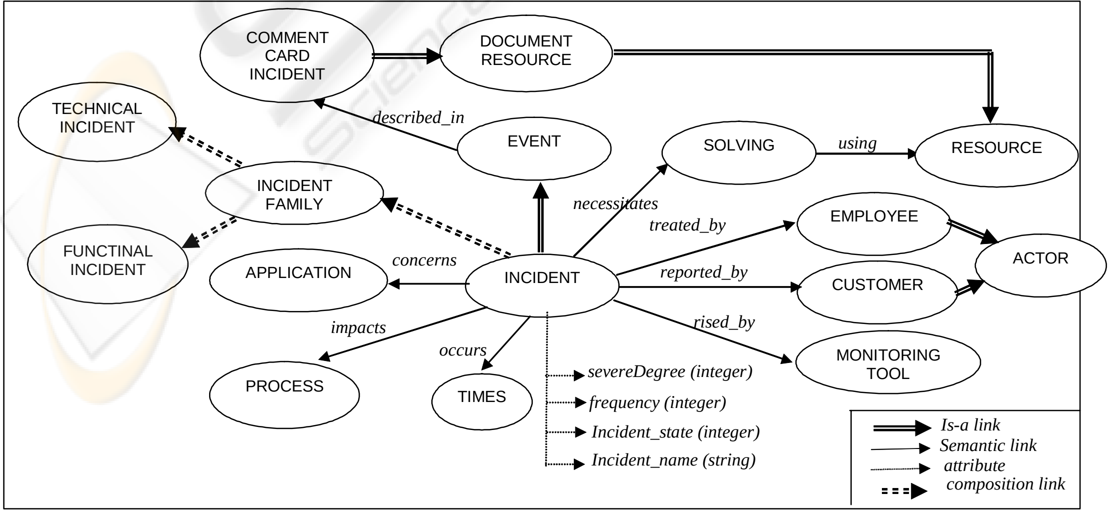

Figure 1: OntoCIRTIL SNF.

OntoCIRTIL is presented in semantic network formalism (SNF). A semantic network is basically a directed graph with nodes representing concepts from the discourse area and edges expressing relationships between these concepts. A concept is defined in a single way. Knowledge relating to a concept is factorized on the level of the node. This knowledge is expressed by attributes (property), and by relationships. For example, attributes of the node *"Incident"* are: *Incident Name*, *Incident State*, *Gravity Degree* and *Frequency*. Relationships linked to *"Incident"* are "*Incident concerns Product" or "Incident treated-by Employee"*.

3 USER'S ENVIRONMENT

Our work contributes to the elaboration of a CM environment. This environment is directed towards several user types: experts of field, actors (end-users of knowledge) and administrators. It allows indexing and integration of knowledge in the CM as well as querying of heterogeneous information sources. These tasks are classified through two processes (capitalization and restitution) which are related with several databases. Hereafter, we describe these tasks and define the role of databases. In this paper, only dark components of Figure 2 are detailed.

The capitalization process consists of the capture, the treatment and the integration of knowledge in the CM. Knowledge is made explicit and reusable. Each actor can participate to this process according to access rights. Main tasks are:

Creation: Achieved directly by actors and can be result of the drafting of business documents, technical reports, actors' experiences, etc. Knowledge must be created or converted regarding the conventions of the company.

Dematerialization: Knowledge dematerialization is an operation that consists of translating knowledge issued from documents or actors experience into a computational form. This allows a huge amount of knowledge to be stored and retrieved.

Formalization: XML and RDF(S) are used for describing syntax and semantics of semi-structured information sources. RDF provides a simple data model for representing formal semantics of information, i.e. meta-information. RDF Schema defines a simple ontology modelling language on top of RDF that can be used to define vocabulary and structure of meta information.

Indexing: Indexing techniques allow to create and to describe objects with their attributes and relationships. Objects are knowledge items found on web pages, in spreadsheets, or in text documents. This task updates the knowledge database (§3.3).

The Restitution process consists of extracting data, documents or document fragments for end-users. Hyper-navigation and interfaces contribute to this process. Two modes of restitution are proposed: navigation through taxonomy of ontology and interrogation based on a search engine.

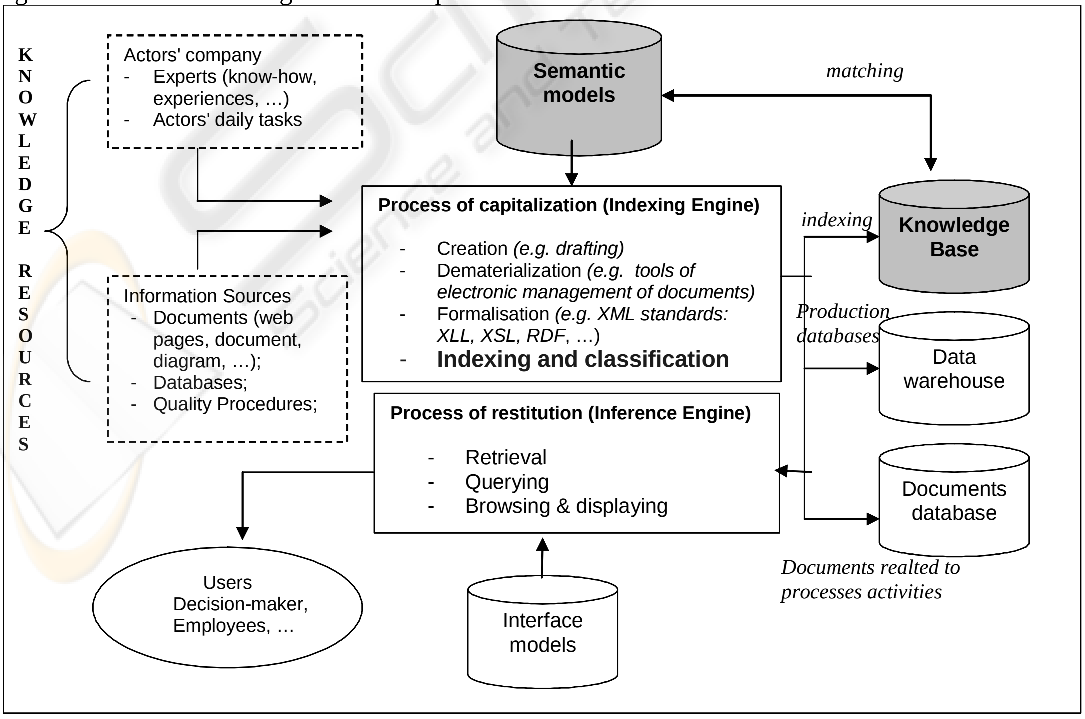
K N O W L E D G E R E S O U R C E S
Actors' company
- Experts (know-how, experiences, …)
- Actors' daily tasks
Information Sources
- Documents (web pages, document, diagram, …);
- Databases;
- Quality Procedures;
Semantic models
matching
Process of capitalization (Indexing Engine)
- Creation (e.g. drafting)
- Dematerialization (e.g. tools of electronic management of documents)
- Formalisation (e.g. XML standards: XLL, XSL, RDF, …)
- Indexing and classification
indexing
Knowledge Base
Production databases
Data warehouse
Documents database
Documents realted to processes activities
Process of restitution (Inference Engine)
- Retrieval
- Querying
- Browsing & displaying
Users
Decision-maker, Employees, …
Interface models

Figure 2: Corporate Memory Environment dedicated to Actors' Company.

Retrieval: Main function allowing users to locate knowledge, search relevant documents. This module is the kernel of inference engine.

Navigation: Actors may explore the CM without asking a particular question. They exploit the ontology for navigation purposes. Also, the ontology allows deriving additional links and descriptions.

Querying: Is a treatment of queries formulated by end-users through a high level language. The query language has to be able to query heterogeneous databases. For instance, we can use XQuery or OQL to query XML-based CM.

Browsing & displaying: Concerns documents (‘fragments) as results of queries and presentation to end-user. User profiles and auto-adaptive interfaces are used for improving access to knowledge.

Roles of databases: A CM is built around many databases. Each database contains data, models and rules that ensure the storage and the treatment of knowledge. We briefly present these databases.

Semantic model: It describes how knowledge is related semantically into the CM. It includes relationships models, properties corresponding to the semantic of each relationship (type of relation, properties of combination with the other relations in the same context) and inferences models. Also, it permits to ensure the coherence supported by required conditions of database management (rules of priority, rules of incompatibility). The semantic model is considered at metalevel which allows cognitive representations of semantics primitives.

Knowledge base: It contains representations (formal) of concepts and relationships of the domain (here: the management of the technical hitches). It contains also the metadata of documentary descriptions (Dublin Core).

Documents base: This database contains structured documents supported by XML standard. The basic structure of an XML-document is given by the hierarchically nested elements, thus it is a natural approach to model it as a tree-graph.

Data warehouse: A data warehouse is a central repository for all or significant parts of the data that an enterprise's various business systems collect. It is enriches from one or more production databases.

Interface models: Several types of models may be useful to represent interfaces and communication setting process. These models can help users to navigate between the nodes of the ontology. A variety of models have been proposed to help retrieve knowledge.

4 THE MODEL S^3

This paper focuses on capitalization process and requirements for a CM. For that we propose an approach to model knowledge aiming to enhance performances of retrieval systems. Before presenting the model S^3, we briefly expose some motivation which guided us to elaborate such a model. Relationships of ontology define and enrich the semantic between the concepts. In certain cases, the type of relationship can change the semantic. Thus, two concepts C1 and C2 linked by two distinct relationships R1 and R2 produce two semantics.

Let us consider two concepts *“Incident”* and *“Actor”* and two relationships *“reported-by”* and “treated-by”. Semantic can be variable according to the relationship employed between each of these two concepts. *Reported-by(Incident,Actor)* means that a given incident has been notified by an actor during his task; and *Treated_by(Incident, Actor)* means that an actor has resolved the incident.

In addition, several types of relationships offer various viewpoints on one concept. In fact, figure 3 shows that the concept *"Incident"* has many significances, defined through relationships. Such explanations are necessary to define all variety semantic of knowledge considered by end-users.

To elaborate such indexing model, we build a domain ontology allowing modelling knowledge according to various facets and viewpoints. Our domain comprises the range of knowledge associated with processes of activity domain of CIRTIL Company. In this section, we present the principles and the components of model S^3.

The *OntoCIRTIL* aims first to represent domain knowledge and then, to be used like a resource of indexing. To achieve this goal, we take advantage of conceptual relationships when we built *OntoCIRTIL.* Although the ontological relationships are numerous, it is difficult to give an exhaustive list of it. This can be explained by the fact that relations in ontologies are not standardized like those in thesaurus. Nevertheless, the model proposed in this paper, called S^3, is based on three views i.e. three spaces for organizing knowledge fragments

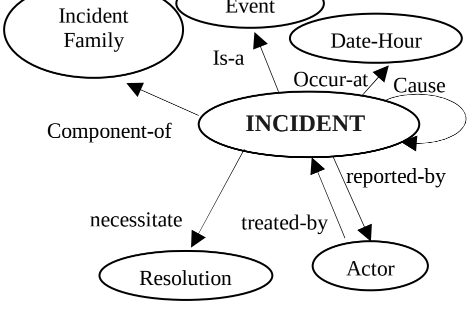

Figure 3: Semantic interpretations of concepts

represented by concept of domain ontology: semantic, structural and subsumption spaces.

Semantic space: This space gives a view allowing discovering concepts across semantic links. This space is domain dependent because relationships used are defined and interpreted by users. Our environment offers primitives to define such relationships such as: create, delete, and rename semantic links. The graphs path algorithms make possible to navigate in this space in order to discover concepts: to displace from a concept to another according to the semantic link defined by the expert. This space embeds *semantic links.*

Structural space: This space takes in account structural dimension of the knowledge concepts and links with document fragments. These fragments are represented by concepts and organized according to aggregation links. Besides the graphs path algorithms, it is interesting to define a zooming function for discovering various levels of the selected concepts. This function allows exploring deeply the various parts of one concept. This space allows representing *structural links*.

Subsumption space: This space makes it possible to organize the definitions of concepts by using subsumption links. This space increases semantic space by using a new dimension which allows improving concepts description in term of related definitions. The organization of knowledge in this space allows actors to retrieve concepts (and related documents) with definitions (which concerns ancestors in the subsumption graph).

4.1 Description of link types of S^3

The concept is the core of our model: concepts are objects (or fragments) of knowledge. Each concept is linked with others concepts by at least a type of links (*Semantic, Structual, Subsumption link).* Each type of relationships gives a particular semantic between two concepts. To model these relationships, our approach is based on manual linguistic analysis. We applied a same method to define the concepts, to choose the representative links, i.e. we proposed several lists which were modified then validated by actors according to the ontological commitment.

Semantic links. The names links are related to the usual language (natural language) of the community for which ontology is available. Several approaches are used for naming links: verbs or prepositions (Sherratt & all, 1990); verbs or nouns (Heeren & all, 1993); verbs or logical connectors (Malone & all, 1984). This facilitates their use in particular to retrieve knowledge in the CM because, these relationships are considered as keywords. Semantic relationships express clearly "evident" and no ambiguous knowledge. Two types of semantic links are illustrated in the following examples.

Example 1: Impact link (Incident, Process)

In this representation the type of semantic link is called *Logical Link.* Generally, the definition of logical link relates to a description of a true logic in the real world. The link between the two concepts: *Incident* and *Process* allows informing the actor (i.e. technician) to take new disposition. Instantiation example is: a server failure as an incident, which impacts the process of exploitation of software located on workstations. Examples of logical links used in our ontology are: *cause, necessitate, implicate,* etc.

Example 2: Occur link (Incident, Data-Hour)

Another type of semantic link is a *temporal link*. In ontology, the concept of time can be even defined at concept level. However, it can be better clarified by a link between two concepts. It is the case of link *"to occur"* established between *Incident* and *Date-Hour*. This representation indicates that any incident can occur at a given moment. Other links such as*: treated-by, written, organized*, etc. compose the relationships employed in our domain ontology. Note that relationships allow the bi-directional and opposite (inverse) semantic expression.

Structural links They express a strong property between the whole and the parts, as well as subordination between the parts and the whole. The parts can be created after the composite itself, but once created, they "lives" and "dies" with him (i.e. they share its duration life). Parts can also be explicitly withdrawn before the death of the composite. Composition can be recursive.

Subsumption links. They are largely used and are considered as the foundation of ontology, because all ontologies are presented in taxonomy. Subsumption allows information about concepts to be associated with their most general concept, and it allows information to filter down to more specific concepts in the taxonomy via inheritance.

In addition our indexing model proposes a mechanism which allows deducting new knowledge in order to enrich a semantic. In some cases, knowledge is also defined with rules*: Chief Project is a person who manages project*. These rules permit to express implicit useful knowledge of information retrieval. We write constraints that control semantic relationships between concepts in order to support reasoning. For example, if *"Incident" necessitates "Resolution"* and if this *"Incident"* is *treated-by "Employee"* then a link is inferred: *"Employee" proposes a solution.*

4.2 Indexing tool

We developed a tool for hitches management called ***MaTIP*** **(*Ma***nagement of the ***T***echnical ***I***ncident ***P***roject*)*. The objective is to capitalize data, information and knowledge allowing the identification, the management and the anticipation of the dysfunction and technical anomalies at the exploitation time of applications. This *Knowledge Base* contains *OntoCIRITIL*. It groups the concepts and relationships identified during the conceptualisation process. The employees can modify, enrich and validate the ontology. One of the objectives of our contribution is to integrate the indexing operation into the daily activities of the actors. To achieve this goal, we take into account that users hardly change their practices.

KnowIndexe is a simple application that actors can use easily to index or to retrieve formalized knowledge from the CM. The indexing technique is achieved through the ontology considered as an indexing resource. The indexing mechanism comprise three steps: *Selection of knowledge* (document or fragment) in the usual environment of actors; *Selection of representative' concepts* in the ontology describing the selected knowledge; *Indexing* in generating a correspondence between concepts and knowledge.

5 CONCLUSION

In this paper we have presented a semantic model based on ontology of domain intended to index technical documents in the context of the CM. We presented first an environment dedicated to actor's company, as a framework for CM development. We outlined the particularities of the domain ontology built for this objective. Then, we presented the model $S^{\mathbf{3}}$ and its components. We explained that the ontological relationships allow a strong semantic. In this context, we proposed three link types. The first experimentation applied to a project of CM permits first, to expose real needs and then to test and validate our approach with the *KnowIndex* indexing tool. The interest of our contribution is to develop an indexing model which exploits the ontological relationships. The application of the model to a small corpus showed that the approach is time-consuming in particular when the ontology must be built. Nevertheless, the implementation of the structural space gave good results for users.

REFERENCES

Benjamins, V. R. and Fensel, D. The ontological engineering initiative (KA)2. In Guarino, N., editor, *Formal Ontology in Information Systems*, pages 287–301. IOS Press.

Desmontils.E, Jacquin.C. "Indexing a Web Site with a Terminology Oriented Ontology", *The emerging semantic web*, I.F. Cruz, S. Decker, J. Euzenat and D. L. McGuinness Ed., IOS Press, Amsterdam, pp. 181-197, 2002, ISBN 1-58603-255-0.

Fernández, M.; Gómez-Pérez, A.; Pazos, J.; Pazos, A. Building a Chemical Ontology using methontology and the Ontology Design. *IEEE Intelligent Systems and their applications*. #4 (1):37-45. 1999.

Gandon.F, Dieng-Kuntz.R, Corby.O, Giboin.A. Semantic Web and Multi-Agents Approach to Corporate Memory Management. *17th IFIP World Computer Congress IIP*, Eds Musen M., Neumann B., Studer R., p. 103-115, August 25-30, 2002, Montréal, Canada.

Guarino, Nicola. 1995. Formal Ontology, Conceptual Analysis and Knowledge Representation, *I. J. of Human-Computer Studies*, 43, 625–640.

Heeren, E. and Collis, B. Design considerations for telecommunications-supported cooperative learning environments: concept mapping as a telecooperation suppport tool. *J. of educational multimedia and hypermedia*, 4(2), 1993, 107-127.

Heflin, J. and Hendler, J. Semantic interoperability on the web. In *Extreme Markup Languages 2000*.

Malone, J. and Dekkers, J. The concept map as an aid to instruction in science and mathematics. *School science and mathematics*, 84(3), 1984, 220-232.

Saadani.L, Bertrand-Gastaldy.S. Conceptual Maps and Thesauri: A Comparison of Two Models of Representation from Different Disciplinary Traditions CAIS 2000. Canadian Association for Information Science *Proceedings of the 28th Annual Conference*.

Sherratt, C. S. and Schlabach, M. L. The application of concept mapping in reference and information services. *RQ*, 30, 1990, 60-69.

Staab S., Schnurr H.-P., Studer R., Sure Y.(2001): Knowledge Processes and Ontologies. *IEEE Intelligent Systems*. 16(1), Special Issue on Knowledge Management, January/February 2001.

Uschold M, Gruninger M. Ontologies: principles, methods and applications. *Knowledge Engineering Review* 1996;11(2).

Van Heijst, G., Schreiber, A.T., Wielinga, B.J. Using explicit ontologies in KBS development. *I. J. of Human-Computer Studies*, 46 (2/3), 1997, pp 183-292.

Weggeman, M. Knowledge management: The modus operandi for a learning organization. In J.F. Schreinemakers, editor, *Knowledge Management—Organization, Competence and Methodology*, pages 175–187. Ergon Verlag,W¨urzburg, D, 1996.